Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 3057–3066

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Explainable Robotics in Human-Robot Interactions

Rossitza Setchi[a,*], Maryam Banitalebi Dehkordi[a] , Juwairiya Siraj Khan[b]

[a]*Research Centre in AI, Robotics and Human-Machine Systems (IROHMS), Cardiff University, Cardiff CF24 3AA, UK*
[b] *Centre for Advanced Studies, Dr. A.P.J.Abdul Kalam Technical University, Lucknow, India*

Abstract

This paper introduces a new research area called Explainable Robotics, which studies explainability in the context of human–robot interactions. The focus is on developing novel computational models, methods and algorithms for generating explanations that allow robots to operate at different levels of autonomy and communicate with humans in a trustworthy and human-friendly way. Individuals may need explanations during human–robot interactions for different reasons, which depend heavily on the context and human users involved. Therefore, the research challenge is identifying what needs to be explained at each level of autonomy and how these issues should be explained to different individuals. The paper presents the case for Explainable Robotics using a scenario involving the provision of medical health care to elderly patients with dementia with the help of technology. The paper highlights the main research challenges of Explainable Robotics. The first challenge is the need for new algorithms for generating explanations that use past experiences, analogies and real-time data to adapt to particular audiences and purposes. The second research challenge is developing novel computational models of situational and learned trust and new algorithms for the real-time sensing of trust. Finally, more research is needed to understand whether trust can be used as a control variable in Explainable Robotics.

Keywords: Explainable Robotics, Explainable AI, Explanation, Reasoning, AI, Robotics.

1. Introduction

Artificial Intelligence (AI) plays an important role in many business models and has become part of the daily lives of everyone who uses smart phones, recommendation systems, online shopping, media streaming channels, e-

* Corresponding author. *E-mail address:* setchi@cf.ac.uk

10.1016/j.procs.2020.09.198

learning platforms, social media or any digital web services. This proliferation of applications has triggered heated discussion among scientists and policy makers regarding trust and explainability: that is, the ability of an AI system to justify the recommendations it makes and explain how it works. A recent study by the Royal Society [1] highlighted a number of concerns regarding accountability, transparency and bias that must be carefully considered when designing AI-enabled systems. The same study argued that different users require different forms of explanations in different contexts [1]. System developers might require technical details about how a system is designed or how data are processed, while an end user may need accessible information about which factors have been considered and how they influenced the decision. Explainable AI is a very challenging area because trust in a system may not always be a desirable outcome. As some recent examples challenging advertisers to explain their strategies on social media have shown, such systems could be used to deceive, mislead or manipulate people.

Similar ethical concerns have been recently expressed with respect to examples of humans trusting robots as much as they trust other humans in emergency evacuation scenarios [2] or of robot automation software making decisions based on biased data, existing patterns of structural discrimination, generalisations or cultural stereotypes. However, although AI is embedded in almost all advanced robotics systems (e.g. robot vacuum cleaners, self-driving cars and social companions), the topic of Explainable Robotics is almost completely absent from discussions about future technologies. A recent study [3] maintained that the lack of ability to provide comprehensive explanations impedes general acceptance of AI and robot systems in critical tasks.

This paper argues that Explainable Robotics is a research area with unique challenges. Different socio-technical contexts require different levels of explainability, as is particularly evident when analysing human-in-the-loop scenarios. For example, patients do not normally question their physicians' authority, but would expect explanations from a robot helping them perform rehabilitation exercises or advising them to continue a specific treatment. Therefore, the new generation of robots fully interacting with humans will require sufficient intelligence to provide explanations in ways that are understandable to humans. The long-term vision is to develop robots that are able to not only explain an action or articulate a response to human partners, but also perform a more active role, in which they participate in a dialogue, debate decisions or suggest second opinions. Explainable Robotics is a small but necessary step towards achieving this vision.

This paper addresses the above challenge by introducing the new research area of Explainable Robotics, which studies explainability in the context of human–robot interactions. The focus is on developing methods for generating explanations that allow robots to communicate with humans in a trustworthy and human-friendly way. The main research challenge is identifying what needs to be explained and how it should be explained by robots to particular individuals.

The paper is organised as follows. The next section reviews the latest research in explainable AI, ethical AI and human–robot interactions. Section 3 introduces Explainable Robotics using a scenario from the healthcare domain. Section 4 discusses the research challenges of Explainable Robotics and highlights the need for new algorithms and further research. Section 5 concludes the paper.

2. Literature review

2.1 Explainable AI

AI systems have become part of our daily lives; one wonders if their decisions and predictions can be trusted. As reported by Pricewaterhouse Coopers (PwC), 82% of interviewed CEOs believe that AI-based systems must provide reasonings and explanations if their predictions and decision-making are to be trusted. AI systems can successfully tackle real-world problems, but their 'black box' nature has affected their application in many safety-critical domains [4] [5].

Explainable AI is a rapidly emerging research area in the machine learning field that attempts to explain how black box decisions are made and how AI models work. It aims to provide ethics, privacy, security, trust, confidence and safety for AI-based systems [6]. These aspects are important for many applications and are increasingly important for personalised medicine and healthcare [7]. Most recently, the European General Data Protection Regulation (GDPR and ISO/IEC 27001) made black box approaches not acceptable in business and the medical

domain, necessitating new models and methods to recreate system decision-making processes and replicate both knowledge extraction and learning. To improve medical diagnosis and pharmaceutical developments, AI decisions must be traceable, and the causality of learned representations must be transparent [5]. The demand in medicine is increasing for AI approaches that are not only accurate, but also trustworthy, reliable, explicit, explainable and interpretable by a human expert [8]. Explanations of AI predictions and decisions are required to achieve thorough trustworthiness of a machine's moral and ethical standards [9]. In the context of explainable AI, there is an important difference between interpretation and explanation: that is, interpretation is the mapping of an abstract notion to an area perceivable by an expert user, while an explanation is a collection of the features of the interpretable area that contributed to the decision making [10]. Several different predictive models have been developed, and each produces a different type of explanation. For instance, an explanation can directly measure the effect of a *local* feature interaction, while different local explanations of each prediction can be combined to interpret the *global* model structure [11].

There are two distinct types of explainability/interpretability: *post-hoc* systems that give local explanations (and not explanations of the whole system's behaviour) for a particular decision to support reproduction on demand, and *ante-hoc* systems, which can be interpreted by design in the direction of glass-box strategies. The model-agnostic framework Black Box Explanations through Transparent Approximations (BETA) is an example of a post-hoc system that can explain any typical black box classifier's behaviour by simultaneously optimising consistency with the original model and the interpretability of the explanation [12]. Decision trees and linear regression models are well-known examples of ante-hoc systems. For example, the decisions of any classifier can be explained visually by formulating them as additive models [13]. This approach has been demonstrated in the context of three different models (i.e. linear support vector machines, naïve Bayes, and logistic regression) [14].

2.2 Ethical AI and Robotics

Ethics in AI and robotics is critical to important issues such as trust, transparency, privacy and security [15]. Ethical deliberation is important because robots typically work in controlled environments in human-in-the-loop scenarios. However, ensuring that a robot works safely and reliably with a human involves several ethical challenges [16] [17]. In an AI-enabled assistive system, human data and privacy protection are essential [18]. Transparency and explainability are also crucial, as transparency in a shared autonomy framework and explainability in AI and robotics will help machines explain their computations, heuristics and decisions. Other challenges include accountability and responsibility, especially in situations in which a robot could cause harm [19].

AI and robotics systems that learn from humans or experience can gradually become smarter, but this poses some serious risks. Trust requires reliability and robustness, but not at the cost of safety and security. In addition, AI systems are vulnerable to cyberattacks, which can jeopardise users' safety. Hence, sufficient measures must be taken to ensure a system's cybersecurity compliance, and user feedback should be sought to improve robots' predictability. In human–robot interaction scenarios in which robots deal with patients (e.g., dementia sufferers), another important challenge is protecting human dignity [17]. Critically, acceptance of both AI and robots must improve significantly [19].

Impartial usage checks should be employed to secure fair and unbiased decisions by AI systems and robots. Internal and external checks should be included in both AI applications and shared autonomy robots to secure egalitarian applications. Impartial AI minimises unfair bias in data and algorithms and minimises human errors during the coding and machine learning stages. Another challenge is avoiding deception. A robot may lead a user to develop convincing but deceptive trust in the system, risking the user's safety and creating a potential threat. Accurate system information may help avoid deception if communicated in advance to the user.

2.3 Human–Robot Interaction

Human–Robot Interaction (HRI) is a field of study dedicated to understanding, designing and evaluating robotic systems for use by or with humans [20]. The concept is applicable not only to robots operated by humans, but also to autonomous systems, as these operate in environments with humans and seek to fulfil human-defined goals and

needs. The type of interaction between humans and robots is highly dependent on whether the robots and humans are in close proximity or separated in space and time. Although the field of Explainable Robotics is also relevant to telemanipulation and software agents, this section reviews only the physical interactions that occur when humans and robots are in close proximity and engage in a dialogue.

Relevant to the concept of Explainable Robotics is Sheridan's classification of autonomy [21], which begins with level 1 (*Computer offers no assistance*) and ends with 10, the highest level of autonomy (*Computer decides everything and acts autonomously, ignoring the human*). Interestingly, all other levels (2 to 9), involve some sort of interaction between robots and humans. For example, at level 2 *(Computer offers a complete set of action alternatives)*, the robot expects a judgement from the human, and at level 5 (*Computer executes the action if human approves)*, the robot needs confirmation before proceeding. Even level 8 (*Computer informs human after automatic execution only if human asks)* may involve human–robot communication when the *if* clause is fulfilled.

Sheridon's levels of autonomy [21] are a good starting point for this discussion on Explainable Robotics. However, more recent studies of autonomy emphasise the collaborative and dynamic nature of robots and humans working together and promote such concepts as shared autonomy [22] and mutual adaptation [23]. In particular, Schilling's three levels of autonomy [22] differentiate among *normative*, *strategic* and *operative* control and could be used to define the types of explanations required at each level.

Beer [24] defined autonomy as the extent to which a robot can *sense* its environment and then *plan* and *act* upon that environment with the intent of reaching a task-specific *goal* (either given or created by the robot) without external control. This definition is very helpful to understand the kinds of explanations a robot may need to provide a human, such as explanations of the environment as it senses it; logical interpretations of this reality; and plans, actions and goals based on external commands and reasoning.

Scholtz's taxonomy [25] of the roles humans and robots can assume in human–robot collaboration tasks (e.g. operator, supervisor, peer, etc.) is particularly relevant to the field of Explainable Robotics because it supports the definition of a broad range of applicable scenarios. For example, in the area of assistive robotics, a robot may act as a mentor to an autistic child or a guide assisting a blind person. In a search-and-rescue scenario, humans and robots could work as peers to secure an area or an unstable structure. All these scenarios require explainability; that is, the robot must be able to explain its actions or recommendations (e.g. justify its choice of a path, detail a particular sequence of steps in a procedure, provide helpful examples). Many of the key challenges in designing such complex human–robot interactions are linked to the need to use real-time information from multiple sources, navigate in complex dynamic environments and, in many cases, interact with vulnerable persons and children. Therefore, one of the key requirements is that the interactions and resulting behaviours can accommodate high levels of complexity [20].

The communication channels used in such complex scenarios combine visual displays, gestures, facial expressions, brain signals, speech, recorded messages, vibrations, recorded alarms and alerts, and more. Some of the exchanges between humans and robots can be scripted and based on a formal language and limited vocabulary, but the more challenging aspects of effective communication require *situational awareness* and understanding of *human cognition and behaviour*.

Situational awareness in the context of human–robot interaction [25] involves perception of cues, an ability to comprehend or interpret them, and an ability to forecast future events and dynamics based on the perception and interpretation of reality. This approach is very relevant to Explainable Robotics, as it highlights the need to build and explain a model of reality, interpret that reality in the context of a specific goal or task, and predict how the reality may change in future. Cognitive modelling plays a very important role in human–robot interactions, as it allows a robot to identify and adjust to a human's cognitive state, while also allowing humans to correctly interpret robots' actions or behaviours. As highlighted by Lemaignan et al. [26], human–robot collaboration supported by multi-modal and situated communication is a challenge for AI, as it involves three aspects: communication, joint action and human-aware execution. Joint action, for example, builds on a common goal, a physical environment and a belief state that includes *a priori* common-sense knowledge and mental models for each of the agents involved (i.e. the robot and its human partners). Addressing such challenges in the context of human–robot collaboration is one of the goals of Explainable Robotics.

2.4 Explanation

People use explanations to improve their understanding of nature, phenomena, situations, events and behaviours.

Effective explanations accommodate novel information in the context of prior beliefs in a way that fosters generalisations [27]. Many important decisions in law, engineering, medicine, politics, diplomacy and everyday life are based on implication-rich, conditionally dependent pieces of evidence, and such decisions often require an explanation-driven approach [28].

An explanation is a set of statements containing facts, beliefs, rules, context, clarifications of causes and possible consequences. There are many different types of explanations because many things in life could be better understood. For example, Aristotle, who first introduced the theory of causality, argued that a 'why' question can be answered by a final, formal, efficient or material cause; therefore, a complete explanation should address four aspects: matter, form, agent and end of purpose [29].

Explainability is a system feature that differs from causability [5] [27]. Many explanations are the result of logical inferences, such as deductive, inductive and, less often, abductive reasoning. Mathematical and philosophical logics are associated with deductive reasoning, science is dominated by inductive reasoning and abductive reasoning is used in areas that use reasoning to develop hypotheses (e.g. medical diagnostics, science). Logical inferences are the foundation blocks of logical AI (e.g. rule-based systems, induction trees). Deep neural networks are interpreted using different approaches. Evaluating *uncertainty* measures prediction and variability (e.g. how model parameters can be affected by small perturbations in training data). *Attribution* methods associate a specific output of a deep neural network to the relevant input variables. Finally, *Activation maximisation* recognises input patterns that resulted in the maximal activations associated to particular classes in the output layer [5].

Formal reasoning is used to support many decision support systems. Relevant to Explainable Robotics is Pennington's model of explanation-based decision making, which views explanation as a mental representation of a situation relevant to a decision and emphasises events, conditions and relationships [28]. Such an approach to explanation can not only improve the quality of the decision making but also provide a requisite explanation if decisions are challenged. Although Pennington's theory is verified in the legal domain, where, according to the authors, explanations are based on spontaneously constructed stories, this approach could serve as a starting point for developing a computational model for Explainable Robotics. In fact, model- and case-based reasoning are two other well-known AI formalisms that support decision traceability.

Recent data-driven AI approaches in robotics demonstrate task performance superior to the symbolic representations used in classical AI, but have made the problem of explainability even more acute. In general, robotic systems lack the ability to explain their own behaviour. There have been some successes in explaining action plans and abnormality detection, but the main problem is the lack of interpretability of the knowledge representations used in data-driven AI. A recent study [3] addressed this problem by linking demonstration, learning, action planning, evaluation and explainability in a framework that uses both symbolic representations and a data-driven haptic model. This approach was successful in generating functional explanations and demonstrated how a specific question (e.g. what is the sequence of actions?) can be addressed in the context of action planning through learning by demonstration (robot explanation: pushing the cap three times and twisting the cap three times). The study demonstrated a significant increase in human trust ratings, but was limited to a specific task and, therefore, lacked the requisite broader perspective: the need to answer the 'why' question.

As previously emphasised, robots need the ability to reason about what needs to be explained and how it should be explained to each individual. The next section discusses these two key aspects using a scenario from the healthcare domain.

3. Explainable Robotics

Figure 1 shows a hypothetical scenario of robot assisting an elderly person with dementia and helping them with activities of daily life and reminiscence therapy. The scenario involves a person with symptoms of early stage of dementia, their friends and family, a caregiver and a robot.

In this scenario, the robot is learning from the caregiver how to assist the elderly person with activities of daily life. The knowledge acquired includes both declarative knowledge (e.g. medication prescribed, medical condition) and procedural knowledge (e.g. how a medication is normally administered) about the person's daily routine. The

robot's knowledge is enhanced through direct observations of interactions and certain procedures (e.g. caregiver offers a glass of water) and involves a transfer of skills from the caregiver to the robot (e.g. how the medication is given or how the glass of water is placed). The robot's knowledge about the person is further enriched through its interactions with the elderly individual and his/her friends and family.

An example of such an interaction is the life story book and its surrounding activities and conversations. The life story book is a form of reminiscence therapy used to help dementia sufferers recall past events and people. In its most simplistic form, the life story book is an annotated photo album with photos and additional artefacts provided by family and friends. The therapeutic effect is achieved through conversations with the elderly individual, which involve extensive reminiscence, associations and helpful reminders. It could be assumed that the robot also has access to lexical resources (e.g. thesauri), knowledge sources (e.g. Wikipedia) and ontological knowledge (e.g. family trees, ontologies of places, professions, historic events, etc.).

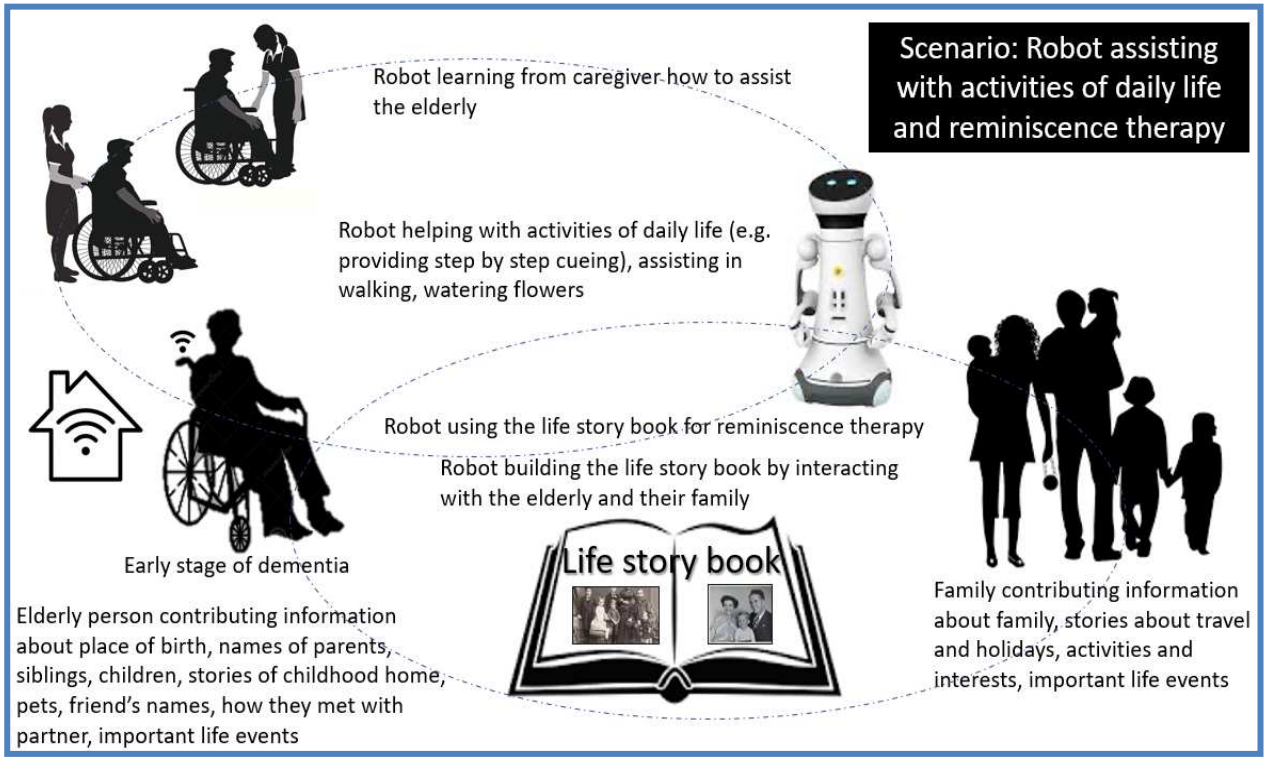

Figure 1. Explainable Robotics scenario of robot assisting with activities of daily life and reminiscence therapy for dementia sufferers

This scenario requires many different types of explanations and very advanced abilities to observe interactions, analyse human behaviour, predict intentions, reason about a person's needs, offer solutions and construct simple dialogues. Extracting meaning from the environment (e.g. how objects are positioned in the house) and behaviours (e.g. gestures, posture, speech, emotions), combining observations with expectations (e.g. comparing how a person sounds with how they normally speak) and reasoning about these facts using background knowledge and context are some of the key abilities required to address the challenges of Explainable Robotics.

Example

Person to Robot (preparing for a visit by a family member, but speaking incoherently and mumbling with jumbled speech; difficult to understand): [incoherent speech]
Robot (observation): *Person is not speaking clearly.*

Robot (reasoning, using knowledge sources and reviewing possible reasons for the cause of the sudden speech problem, such as tiredness, stress, too much to drink, stroke, migraine, neurological disorder, or medication; exploring all these possibilities, but not finding a convincing cause because the medication has been already taken and the person is smiling and looks in good health): *Person looks ok.*

Robot (engaging in more observation, reasoning, and building environmental awareness, notices an unusual object on the table; uses knowledge sources to identify the object as dentures): *There are dentures on the table. These dentures belong to Person.*

Robot (finding a solution, realising that it needs to offer help as the person will be meeting his/her family shortly): *Person needs dentures*.

Robot to Person (explaining/asking permission to act by adapting the explanation and speaking slowly, with simple vocabulary and sentence structure): *Your dentures are on the table. Do you want them?*

Person to Robot (confused expression, not realising they need the dentures): [incoherent speech]

Robot to Person (offering a simple explanation): *You need your dentures to talk with your family.*

Robot to Person/Caregiver/Family (offering a more detailed explanation): *Person did not speak clearly. Person looked ok. There were dentures on the table. These dentures belonged to Person. Person needed dentures to talk with family.*

In this example, the human's speech is different from his/her normal speech, and the human has particular difficulties with labiodental sounds (i.e. sounds that involve the teeth and lips). The robot explores the possible causes of sudden speech problems and assigns a low probability to each based on additional observations of the human. The robot expands its search space by exploring the environment and looking for deviations from the normal morning routine in the bedroom. The unusual object found on the table is identified as a set of dentures. Through continuous observation, the robot formalises the instances of daily life as cases, which include features characterising human and features characterising the environment. In this specific instance, the unusual elements in the case are 'speech' and 'dentures'. Using common sense knowledge, the robot finds a causality between the two, evaluates the situation and decides to offer help. The human looks confused, so the robot explains *why* the human may need his/her dentures. The robot could also provide a longer explanation following the logical chain of reasoning, which could be offered to all actors in the scenario; however, the simpler explanation is deemed more suitable for talking with the dementia sufferer.

Clearly, in the context of robotics, trust and autonomy, explanations are *primarily* needed to explain why a robot has undertaken or is planning to undertake a particular action. There are many types of explanations (e.g. deductive-nomological, historical, psychological, functional, pragmatic, etc.) [27] [30], and more research is needed to understand how to construct human-friendly and trustworthy explanations using analogies, past experiences and real-time information.

It should be noted that although in human–robot interactions, trust may not always be a desirable outcome, as a system could be used to deceive, mislead or manipulate people, quantifying trust in real time requires further research, at a minimum, scenario-based metrics can be defined for each scenario. For example, in health care scenarios, trust could be measured through emotional feedback, gestures, physical and visual contact and willingness to cooperate. Similar to assessing accuracy levels in the execution of an algorithm, Explainable Robotics

needs a mechanism to assess levels of trust and acceptance and identify possible manipulations, hackings and unexpected learned behaviours.

4. Research Challenges

Explainable Robotics is critical for the successful deployment of many applications that involve humans. However, as a new research area that has inherited many of the unsolved problems of both AI and robotics, it faces two significant challenges.

The first challenge is the need for new algorithms for generating explanations that use past experiences, analogies and real-time data to adapt to a particular audience and a particular purpose. A good starting point to determine what needs to be explained is Sheridan's classification of autonomy [21], which differentiates among levels of autonomy. As stated before, all but two of the ten levels of autonomy involve human interaction. Different types of explanations can be offered at different levels as the purpose of the explanation changes (e.g. a set of alternatives at level 2, a tentative decision requiring approval at level 5 and a retrospective account of actions at level 8). Explanations of results and strategies used to achieve an outcome could be based on past experiences, data modelling, feedback, historical facts or interpretations of events, and robots should be able to choose the correct method for each explanation. Past experiences and acquired knowledge provide explanations with examples, analogies and means to compare and rank alternatives. In cases in which robots use data to provide decision makers options, explanations may contain information about variables, data biases, the number of options considered, confidence levels and use of weights/probabilities. Explanations based on risk propensity could consider harm, defects, biases and feedback from previous decisions. Robots should be able to explain their own learning processes and demonstrate how they have reached specific decisions in terms of learning strategy (e.g. criteria used to discard an option), rate of learning (e.g. how and why the model has been updated), data sources and quality (e.g. completeness, credibility, comparability, consistency), deductions and inferences used and type of feedback mechanism employed. Explaining security vulnerabilities may involve references to security access levels and requirements for traceability.

The second research challenge is the need for novel computational models of situational and learned trust and new algorithms for real-time trust sensing. Achieving human-like trust in human–robot collaboration scenarios poses enormous difficulties and hurdles, but these are outweighed by the potential benefits. A durable and trusted working relationship is critical when seeking to understand our surroundings, evaluate hazardous situations, demonstrate empathy and understand human or robot behaviours. Understanding why a robot does an action can help ensure robots are working as expected, systems comply with regulatory standards and any unexpected or newly learned behaviours can be understood by human collaborators. Real-time sensing of trust requires improved algorithms for detecting attention and acceptance, understanding gestures and body language and determining the effectiveness of the selected explanation type in the given context. Finally, more research is needed to understand whether trust can be used as a control variable in Explainable Robotics.

5. Conclusions

This paper introduces the new research area of Explainable Robotics, which studies explainability in the context of human–robot interactions. Explainable Robotics offers many research challenges and has the potential to enhance numerous robotic applications, with significant and positive social and economic impact.

This research area focuses on developing algorithms to provide explanations that allow robots to communicate with humans in a trustworthy and human-friendly way. The research challenges include identifying what needs to be explained in a human–robot interaction scenario, determining how the robot should explain the situation to an individual and discovering how to best use machine learning approaches to enhance these explanations.

Acknowledgements

The authors would like to acknowledge the support of the Centre for Artificial Intelligence, Robotics and Human-Machine Systems (IROHMS) operation C82092, partially funded by the European Regional Development Fund (ERDF) through the Welsh Government.

References

[1] Royal Society (2019). Explainable AI: The Basics. https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf.

[2] Robinette, P. , Li, W., Allen, R., Howard, A.M., Wagner, A.R. (2016). Overtrust of robots in emergency evacuation scenarios. 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7-10 March 2016, pp. 101-108.

[3] Edmonds, M. Gao, F. Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y. (2019). A tale of two explanations: enhancing human trust by explaining robot behavior. Science Robotics, 4(37).

[4] Core, M.G., Lane, H. C. , Van Lent, M., Gamboc, D. (2006). Building explainable artificial intelligence systems. MIT Press, AAAI, pp. 1766–1773.

[5] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

[6] Kiesenberg, P., Weippl E. W., Holzinger, A. (2016). Trust for the doctor-in-the-loop. European Research Consortium for Informatics and Mathematics (ERCIM) News: Tackling Big Data in the Life Sciences. 104(1), pp. 32–33.

[7] Hamburg, M. A., Collins, F. (2010). The path to personalized medicine. New England Journal of Medicine, 363(4), pp. 301–304.

[8] Hudec, M., Bednarova, E., & Holzinger, A. (2018). Augmenting statistical data dissemination by short quantified sentences of natural language. Journal of Official Statistics (JOS), 34, pp. 981–1010

[9] Skirpan, M., Yeh, T. (2017). Designing a moral compass for the future of computer vision using speculative analysis. Proc. of IEEE Computer Vision and Pattern Recognition, pp. 64-73.

[10] Montavon, G. , Samek, W., Muller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, pp. 1-15.

[11] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 2522-5839.

[12] Schoenborn, J. M., Althoff, K.-D. (2019). Recent Trends in XAI: A broad overview on current approaches, methodologies and interactions. In Case-Based Reasoning for the Explanation of Intelligent Systems (XCBR) Workshop, pp. 1-10.

[13] Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., MacDonell, C., Anvik, J. (2006). Visual explanation of evidence with additive classifiers. In National Conference on Artificial Intelligence, AAAI. pages 1822-1829.

[14] Szafron, D., Lu, P., Greiner, R., Wishart, D. S., Poulin, B., Eisner, R., Lu, Z., Anvik, J., Macdonnell, Fyshe, A., Meeuwis, D. (2004). Proteome analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. Nucleic Acids Research, 32(S2): W365-W371.

[15] Veruggio, G., Operto, F. (2006). Roboethics: a bottom-up interdisciplinary discourse in the field of ap-plied ethics in robotics. International Review of Information Ethics Vol. 6.

[16] Deng, B. (2015). The Robot's Dilemma. Nature, Vol 523, 2 July 2015.

[17] Körtner., T. (2016). Ethical challenges in the use of social service robots for elderly people. Zeitschrift für Gerontologie und Geriatrie 49, pp. 303-307.

[18] Sharkey, N., Sharkey, A. (2012). Robotic surgery and ethical challenges. Medical Robotics- Minimally Invasive Surgery, pp. 276-291.

[19] Hameed, I. A., Tan, Z.-H., Thomsen, N., Duan, X. (2016). User acceptance of social robots. ACHI 2016: The Ninth International Conference on Advances in Computer-Human Interactions. pp. 274-279.

[20] Goodrich, M. A., Schultz, A. C. (2007). Human-robot interaction: a survey. Foundations and trends in human-computer interaction, 1(3), pp. 203-275.

[21] Sheridan, T. B. (2002). Humans and Automation: System Design and Research Issues. John Wiley and Sons.

[22] Schilling, M., Kopp, S., Wachsmuth, S., Wrede, B., Ritter, H., Brox, T., Nebel, B., Burgard, W. (2016). Towards a multidimensional perspective on shared autonomy. In 2016 AAAI Fall Symposium Series.

[23] Nikolaidis, S., Zhu, Y. X, Hsu, D., Srinivasa, S. (2017). Human-robot mutual adaptation in shared autonomy. Proc ACM SIGCHI. Mar; 2017, pp. 294–302.

[24] Beer, J. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. J Hum Robot Interact. 3(2), pp. 74–99.

[25] Scholtz, J. (2003). Theory and evaluation of human robot interactions. In Proc. Hawaii International Conference on System Science 36.

[26] Lemaignan, S., Warnier, M., Sisbot, E.A., Clodic, A., Alami, R. (2017). Artificial cognition for social human-robot interaction: An implementation. Artif. Intell. 247, pp. 45-69.

[27] Lombrozo, T. (2006). The structure and function of explanations. TRENDS in Cognitive Sciences, 10(10), pp/ 464-470.
[28] Pennington, N., Hastie, R. (1993). Reasoning in explanation-based decision making. Cognition, 49, pp. 123-163.
[29] Falcon, A. (2019). Aristotle on Causality. in The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), Stanford University, USA.
[30] Wilkinson, S. (2014). Levels and kinds of explanation: lessons from neuropsychiatry. Frontiers of Philosophy, vol. 5, Article 373.